Education and Science

Vol 42 (2017) No 189 217-247

An Analysis of Rater Severity and Leniency in Open-ended Mathematic Questions Rated Through Standard Rubrics and Rubrics Based on the SOLO Taxonomy *

Bayram Çetin [1], Mustafa İlhan [2]

Abstract

This study uses the many-facet Rasch model to analyze the severity and leniency of raters of open-ended mathematic questions rated through standard rubrics and rubrics based on the SOLO taxonomy. The data source of the study was obtained from 104 eight grade students based on their responses to open-ended questions on mathematics-achievement test that was created by the researchers. The study's participants were seven mathematics teachers who serving as raters in the research. The data collection instruments involved standard rubrics and rubrics based on the SOLO taxonomy. The collection of the data was performed in a few phases. In the first phase, the mathematics achievement test, which included open-ended questions, was administered to the students and evaluated by the raters. Then, raters scored students' responses to the open-ended questions using standard rubrics. Next, the raters scored the responses using rubrics based on the SOLO taxonomy. The data thus acquired were analyzed using the many-facet Rasch model. The study found that the raters' agreement was low, and that there were significant differences between raters in terms of severity and leniency when they used standard rubrics. When the ratings were done with rubrics based on SOLO taxonomy, the raters were consistent and their scoring severity and leniency were similar to each other.

Keywords

Open-ended questions
Severity and leniency of raters
Rubric
SOLO taxonomy
Many facet Rasch model

Article Info

Received: 07.29.2015
Accepted: 11.07.2016
Online Published: 02.21.2017

DOI: 10.15390/EB.2017.5082

* This study includes a part of the PhD thesis by Mustafa İLHAN.

[1] Gazi University, Gazi Faculty of Education, Department of Educational Sciences, Turkey, bcetin27@gmail.com

[2] Dicle University, Ziya Gökalp Faculty of Education, Department of Basic Education, Turkey, mustafailhan21@gmail.com

Introduction

Evaluation of Mathematical Achievement

It is not possible to determine students' actual levels of learning mathematics subjects since achievement is an abstract phenomenon that cannot be observed directly, but only measured indirectly (Tan, 2015). In arguments about a student's level of learning a subject, the student's performance of a task or responses to questions play an essential role. Thus, studies of the evaluation of the mathematical performance are based on the assumption that students' responses to items on a valid and reliable test are a valid indicator the level of their abilities. This makes the selection of the questions and tasks for the evaluation of students' mathematical achievement a critical component of the assessment process (Romberg & Wilson, 1992). The methods used in the mathematical assessment process must be capable of measuring the skills students attain in mathematical classes. These skills are: *i)* comprehension of mathematical concepts and systems, *ii)* using these concepts and systems in real life and in other fields of learning, *iii)* using mathematical terminology correctly to explain personal opinions, *iv)* creating arguments using induction and deduction, *v)* developing problem-solving strategies and applying them to problems in everyday life (Ministry of National Education [MNE], 2009; National Council of Teachers of Mathematics [NCTM], 2000). Multiple-choice tests are not sufficient for the measurement of these skills. Thus, methods beyond multiple-choice tests are needed in mathematics evaluation process. This gap is filled by the natural aspects of performance assessment (Güler, 2008).

Performance Assessment

Performance assessment has been described variously by many researchers. Since researchers assign different meanings to performance assessment, it is difficult to delineate this concept (Palm, 2008). Stecher (2010) suggested focusing on what performance assessment is not, rather than what it is to be able to determine its limits clearly. Performance assessment is not a multiple-choice, true-false or matching test (Stecher, 2010). Instead, students should respond for themselves in performance assessment (Zhu, 2009). This requires students to structure the information instead of merely retrieving it from their memories as they do in multiple-choice tests (Moore, 2009). In other words, performance assessment requires students to solve complicated problems, show their work (McBee & Barnes, 2009) and justify their responses (Woodward, Monroe, & Baxter, 2001). These aspects of performance assessment help students to see their strengths and weaknesses and gives them more detailed information about the things they learn. They also participate in the learning process more actively, explain their thoughts freely, use their mathematical knowledge and mathematical thinking skills and build relationships between the things they learn (National Assessment Governing Board [NAGB], 2002). All of these points help students improve their superior cognitive skills (Kind, 1999). Thus, performance assessment is a better way of measuring the complicated skills and communicative competencies required by modern societies (Palm, 2008).

Performance assessment has several limitations in addition to its listed advantages. The primary limitation of performance assessment is that it cannot be scored objectively like multiple-choice tests (Romagnano, 2001). Students' scores on a test which is not scored objectively may vary according to the rater (Tekin, 2009). In the literature there are several studies exemplifying this situation (Özmantar, Bingölbali, & Akkoç, 2008; Güler, 2008; Kan, 2005; Koretz, McCaffrey, Klein, Bell, & Stecher, 1992; Toffoli, Andrade, & Bornia, 2016). For instance, in a study by Koretz et al. (1992), students' mathematical performances were scored by two raters using a rubric with four points and found that the consistency between the two raters was weak. In the study by Özmantar et al. (2008), 171 teachers scored the same student's responses to an open-ended mathematical question and gave different scores ranging between 0 and 10. Of the teachers, 44% gave 10 out of 10 points while 24% gave 0 points for the same response.

Another study by Bingölbali, Özmantar, and Akkoç (2008) found that a majority of teachers gave privilege to practical solutions based on rules and disregarded different solution methods while scoring students' responses to open-ended mathematical problems. A study conducted by Güler (2008) could also stand as an example for rater differences in performance assessment. In the research conducted by Güler (2008) students' responses to open-ended math questions were scored by four different raters, and the results were evaluated using a many-facet Rasch model. The Rasch analysis indicated that the consistency among the raters was weak, and they were inclined to give different scores for the same response. This shows that students' performances on open-ended questions do not only depend on their ability levels, but were also influenced by factors caused by the raters (e.g., raters' age, gender, scoring experience, previous training in assessment). The rater's influence on students' performances is called *rater effect* (Farrokhi, Esfandiari, & Vaez Dalili, 2011). Since rater effects lead to variance that is not related to the criteria used to assess students' test scores (Eckes, 2005; Hoyt, 2000), these effects increase the errors in the assessment and reduce the reliability of judgements about students' ability levels.

Rater effects in performance assessment include raters' leniency and severity, central tendency, halo effect, range restriction (Saal, Downey, & Lahey, 1980) rater bias (differential rater leniency and severity) and inconsistency (Myford & Wolfe, 2004). However, Cronbach (1990) said that raters' leniency and severity are the most important rater effect in performance assessment. Raters' leniency and severity refers to raters' regularly inclination to give higher or lower scores than other raters or benchmark ratings (Jackson, Schuler, & Werner, 2009). Rater's leniency and severity can lead to a student rated by a severe rater obtaining a lower score than other students with weaker skills who are graded by a lenient rater (Wiseman, 2012). This tendency reduces the consistency between raters. The consistency in the scores given to the same person by the two or more raters is accepted as the criterion for reliability between the raters (Moskal & Leydens, 2000). Thus, multiple raters giving scores with different severity and leniency means that the reliability between the raters is low, while it shows high reliability between raters when they give scores with similar severity and leniency.

The severity and leniency of raters may cause students, who are rated by severe raters, to obtain lower scores than other students who are less talented at the subject of the assessment, but evaluated by more lenient raters (Wiseman, 2012). In this situation, the variance between the scores of the students reflects not only their competency levels, but also includes the raters' severity and leniency. Rater severity thus has serious implications, particularly for students who have score near the cut score (whose ability levels correspond to the criterion used in the assessment), and these results may be difficult to compensate for. For instance, if a senior student is scored by a severe rater, that student may have to continue school for another semester or even an entire academic year (McNamara, 1996). On the other hand, rater leniency can help students who fail to achieve their goals to pass a course or graduate. For instance, when medical students' suturing skills are assessed by a lenient rater, it can lead to inaccurate evaluations of students who have insufficient suturing skills. Thus, the severity and leniency of the raters should be minimized to obtain accurate results. Raters should use a common approach to the criteria to be used in assessment to control rater severity and leniency. Rating students' responses using rubrics is one of the most important ways to establish such a common approach.

Standard Rubrics and Rubrics Based on the SOLO (Structure of Observed Learning Outcome) Taxonomy

Rubrics are guides for rating that describe the characteristics and criteria for different levels of performance (Kan, 2007). Rubrics both increase the consistency between the scores of raters who score the same performance and prevent one rater from giving different scores to the same performance on different occasions. Thus, rubrics make it possible for rating to be performed regardless of when or by whom it is done (Moskal & Leydens, 2000). Some rubrics are created with no supporting taxonomy, and in this study, they are called *standard rubrics*. In standard rubrics, the criteria for evaluation are determined without any taxonomy that supports them and the ratings are done by using different degrees such as "*inadequate*," "*developing*," "*acceptable*," "*good*" and "*very good*" (Gronlund, 1998). The rater identifies the level of the rubric suited to the student's response considering the steps in solving the problem, the correctness of the result and the sufficiency and comprehensibility of the explanations of the solution. For instance in a standard rubric, if a student's response and solution to an open-ended mathematical problem are both incorrect, the level is "*inadequate*." If the response is correct, but the operations and the solution are incorrect, the level is "*developing*." If the strategies used in the solution are correct and clear but the response is incorrect because of few mistakes, the level is "*good*". If the response is correct, and the steps of the solution are clear and understandable, then the level is "*very good*." With the aim of presenting an example for standard rubrics, Table 1 illustrates a holistic rubric that is included in the teachers' guide published by the Ministry of National Education (MNE) (2007) and addressed to measure students' problem solving skills in mathematics.

Table 1. Examples of Standard Rubrics for Mathematics Problem Solving Skills

Criteria		**Score**
1 Point	If the solution has the following characteristics, this point will be given	
	-No operations have been made -The student wrote only the incorrect answer -The students only copied the data in the problem or there are no signs of understanding the problem	
2 Points	If the solution has the following characteristics, this point will be given	
	-The student operated only on one of the subgoals and has not reached any results -The student started to make operations to solve the problem, yet the operations were not sufficient to find the correct answer -The student started with an inappropriate strategy or tried to solve the problem using that strategy and failed	
3 Points	If the solution has the following characteristics, this point will be given	
	-The student understood the problem but reached an incorrect result since her or she started with an incorrect strategy -The student found the correct answer but the solution is not clear -The correct answer is provided but without any operations -The only correct answer belongs to one of the subgoals of the problem -Only the beginning of the problem is solved with the correct strategy -The student selected the correct strategy but used it incorrectly	
4 Points	If the solution has the following characteristics, this point will be given	
	-The student found an incorrect answer since he or she partly understood the problem or did not understand at all -The student reached an incorrect answer due to unclear reasons despite using an appropriate strategy -The student gave a correct answer, although the correct strategy is not observed in the operations -The student used the correct strategy but did not write the result	
5 Points	If the solution has the following characteristics, this point will be given	
	-The student made an error using the correct strategy, but this error resulted neither from misunderstanding the problem nor from a misconception -The student used the correct strategy and reached the correct solution	

Some researchers use the reflective thinking model, Bloom taxonomy or SOLO taxonomy to determine the assessment criteria for rubrics (Chan, Tsui, Mandy, & Hong, 2002). Rubrics based on the SOLO taxonomy are often used to rate open-ended questions at various educational levels and in various courses (Hattie & Purdie, 1998). The SOLO taxonomy was created by Biggs and Collis (1982) with the purpose of explaining the structure of the observed learning outcomes. In the SOLO taxonomy, the learning cycle has five different levels. They are: prestructural, unistructural, multistructural, relational and extended abstract (Mohd Nor & Idris, 2010). These levels represent five different ways students can structure their responses to any kind of question (Lucas & Mladenovic, 2008). On the prestructural level, the student fails to perform the task appropriately. The student's arguments do not provide any help to solve the problem (Leung, 2000). The answer provided by the student is not relevant to the problem itself (Brabrand & Dahl, 2009). On the unistructural level, the student approaches the subject with a shallow and limited viewpoint and focuses on a single aspect of it. On the multistructural level, the student understands multiple aspects of the subject, yet fails to build a correlation between them. Students' explanations of the solution and the opinions they express include many components. However, the organization of the opinions is poor. The student cannot put the opinions he or she produced together in a consistent way (Leung, 2000). On the relational level, the student sees different sides of the problem and manages to integrate them. On this level, concepts are applied to similar situations or problems. However, they cannot be transferred to another field (Kanuka, 2011). Finally, on the extended abstract, the student can perform reflection and evaluation, create hypotheses and transfer their learning to another field using inductive, deductive and combinational thought processes (Lake, 1999).

The Advantages of Rubrics Based on the SOLO Taxonomy

There are many advantages of using rubrics based on SOLO taxonomy to evaluate students' performance. The SOLO taxonomy helps to determine students' deficiencies in the learning process and makes it possible to do partial credit. This aspect of the SOLO taxonomy makes it suitable for formative assessment, which aims to determine students' strengths and weaknesses and eliminate deficiencies and mistakes. It is also suitable for summative assessment (Hattie & Purdie, 1998). Another strength of rubrics based on the SOLO taxonomy is that they identify both qualitative (deeper learning) and quantitative (superficial learning) aspects of learning (Burnett, 1999). Students who do not understand or misunderstand a subject in the prestructural level of SOLO taxonomy focus on merely a single side of the subject on the unistructural level and can list many aspects of the subject without building correlations between them on the multistructural level. Thus, there is a quantitative increase in students' learning from the prestructural level to the multistructural level (Rembach & Dison, 2016). On the relational level, students can make a meaningful integration of the aspects they list in the multistructural level. Then, on the extended abstract level, they can transfer this meaningfully integrated knowledge to another field. The relational and extended abstract levels reflect the qualitative side of learning (Brabrand & Dahl, 2009). The rubrics based on SOLO taxonomy are used in a variety of lessons to rate open-ended questions due to these advantages. One of these lessons is mathematics (Collis & Romberg, 1992; Lian & Yew, 2012). Although the SOLO taxonomy was not created specifically for the assessment of mathematical outcomes, its levels parallel different forms of mathematical thinking such as algebra, statistics and geometry (Jurdak, 1991; Lian & Idris, 2006; Mooney, 2002). Thus, rubrics based on the SOLO taxonomy are commonly used to rate open-ended mathematical questions. Despite their common use, there are no studies in the relevant literature examining the effectiveness of rubrics based on SOLO taxonomy in controlling rater effect.

The Objective and Importance of the Study

This study uses the many-facet Rasch model to analyze the severity and leniency of raters of open-ended mathematical problems using standard rubrics and rubrics based on the SOLO taxonomy. The study findings will identify which of the two rubrics that were created to score open-ended mathematical questions is more effective in eliminating the differences between raters and improving rater reliability. Moreover, they can indirectly guide assessment studies in other disciplines since the SOLO taxonomy is not dependent on content (Kanuka, 2011). The MNE and Student Selection and Placement Center (SSPC) plan to include open-ended questions into the national examinations in upcoming years, which implies that this study will make great contributions to large-scale testing in the future. In the statement made by MNE regarding the secondary school entrance exam, it was claimed that when FATİH project was initiated with all its components, each student would have a tablet, which could enable open-ended questions to be included in the large scale assessments (MNE, 2013). Similarly, SSPC initiated the Open-Ended Test Project as of 2013 (Student Selection and Placement Center [SSPC], 2013). In framework of this project, it is suggested that open-ended questions will be included in the central examinations along with multiple choice questions, particularly on examinations with fewer candidates. Thus, it is expected that the findings of this study will also be useful for large-scale tests and examinations.

Although there is plenty of theoretical knowledge in the relevant literature about the fact that rater reliability should be high in assessments that use rubrics based on SOLO taxonomy, there are few empirical studies of the effect of these rubrics on rater reliability (Burnett, 1999; Chan et al., 2002; Hattie & Purdie, 1998). These studies derived contradictory findings about how the rubrics based on SOLO taxonomy affected rater reliability. For instance, Hattie and Purdie (1998), Burnett (1999) and Chan et al. (2002) reported that rubrics based on the SOLO taxonomy increase the rater reliability. On the other hand, Leung (2000) and Chan, Hong, and Chan (2001) claim that these rubrics achieve low rater consistency. These different results show that there is a need for new research on this subject. The authors believe that this study will meet this need and thus contribute to the relevant literature.

To determine accurately how rubrics based on the SOLO taxonomy influence rater reliability, future studies about this subject should be planned carefully. Studies in the relevant literature examine the influence of rubrics based on the SOLO taxonomy without comparing them to standard rubrics. The effect of the SOLO taxonomy on rater reliability was attempted to be identified using Bloom's taxonomy by Hattie and Purdie (1998), using a scoring key developed by the raters themselves in the research by Çetin, Boran, and Yazıcı (2014), and comparing the SOLO taxonomy with the rubrics based on Bloom's taxonomy and the reflective thinking model by Chan et al. (2002). All types of rubrics are supposed to reduce differences between raters and increase their consistency (Airasian, 2005). Thus, it is necessary to do a comparative analysis of the ratings done with these two types of rubrics in order to determine whether the rubrics based on SOLO taxonomy are more influential in eliminating the difference between raters and increasing rater reliability than the standard rubrics which are created without any supporting taxonomy. If this comparison is not made, it is not possible to determine whether the influence of rubrics based on SOLO taxonomy on rater reliability resulted from the use of rubrics in the rating process or from the fact that the SOLO taxonomy is the basis of the rubrics used for the rating.

In the studies on determining how rubrics based on the SOLO taxonomy influence rater reliability, the method used to examine reliability is very important. A review of the relevant literature indicates that the studies exploring the influence of rubrics based on the SOLO taxonomy on rater reliability (Burnett, 1999; Hundzynski, 2008; Leung, 2000; Yazıcı, 2013) use the techniques which depend on classical test theory such as the correlation coefficient between raters, simple percent agreement and comparing the means of the raters. This study will use the many-facet Rasch model. According to the many-facet Rasch model, the factors that can affect students' test scores may not be limited to their ability levels or the difficulty levels of the items used in the assessment when the assessment is performed using open-ended questions and factors related to raters might also cause differences in students' test scores (Baird, Hayes, Johnson, Johnson, & Lamprianou, 2013). This aspect of the Rasch model makes it suitable for open-ended questions that are designed in a subjective way (Mulqueen, Baker, & Dismukes, 2000). The many-facet Rasch model is stronger than the classical test theory in psychometric grounds since it is capable of considering multiple error sources simultaneously, determining the interactions between different error sources (Haiyang, 2010), generating ability estimations with higher validity (İlhan, 2016) and providing information at individual level about the persons whose performance is evaluated, the raters and the items instead of providing it at the group level (Barkaoui, 2008). For these reasons, this study uses the many-facet Rasch model, which is different from previous studies that examined rater reliability using methods based on classical test theory. This is another aspect of this study which will contribute to the relevant literature.

Method

This section explains the study's data sources, participants, data collection tools, implementation and the statistical methods used for data analysis.

Data Source

The data source of the study is the responses to open-ended questions on an achievement test created by the researchers given by 104 eight grade students (46 females and 58 males). The achievement test administered to the students was created for a mathematics course. To create the test, the authors initially prepared 18 open-ended questions about numbers, geometry, algebra, measuring, probability, and statistics. After the items were written, the authors sought the opinions of ten experts to evaluate the understandability and appropriateness for eighth grade students. Table 2 presents the demographic data of these experts. The experts assessed the items using a three-degree scale that said, *The item can be included in the measurement tool as it is* (3), *The item can be included in the measurement tool after it is revised* (2) and *The item should be excluded from the measurement tool* (1). This study is not concerned with determining students' achievement in any mathematical subject. Thus, the authors did not do any analyses regarding content validity. Based on the opinions of the experts, the authors deleted six items that were thought to be inappropriate for eighth grade students or might not be understandable enough, and some items were expressed in more understandable ways.

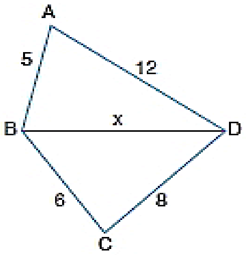
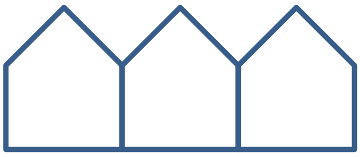
After consulting the expert opinions, the final form of the test included 12 questions since six items were deleted and five items were revised. The authors performed a pilot study with a small student group before finalizing the test and the authors administered the test to 13 eighth grade students; of them, seven were females and six were males. The aim of this pilot practice was to learn students' opinions about the test items and the directions given at the beginning of the test. The authors paid special attention that the student group with whom the pilot practice was conducted represented the lower, moderate, and higher achievement levels. After the pilot practice, the authors decided on ten questions which did not have any problems of understandability. However, two items were deleted since a ten-question test might be problematic in terms of duration. Then, the authors conducted another pilot practice with 15 eighth grade students (seven females and eight males) in order to receive feedback about the duration of the eight-question achievement test and review the items regarding their understandability. During the pilot practice, there were no statements by students expressing that they could not understand the directions at the beginning, nor were there statements expressing that any of the test items were unclear. Considering the durations of the student who finished the test the earliest and that of the one who finished the test the latest in the preliminary practice group, the authors decided the test duration would be 40 minutes.

Table 2. Demographic Data of Experts Who Evaluated Open-Ended Questions for Understandability And Appropriateness for Eighth Grade Students

Experts	Gender	Educational Status
1	Man	He is an associate professor in mathematics education.
2	Man	He is an associate professor in mathematics education.
3	Woman	She is a graduate in Elementary Mathematics Teacher Education. She also has an M.A. degree in Mathematics Education and is a Ph.D. student in Curriculum and Instruction.
4	Man	She is a graduate in Elementary Mathematics Teacher Education. He also has an M.A. degree in Curriculum and Instruction and is a Ph.D. student in the same field.
5	Woman	She is a graduate in Elementary Mathematics Teacher Education. She also has an M.A. degree in Curriculum and Instruction and is a Ph.D. student in the same field.
6	Man	She is a graduate in Elementary Mathematics Teacher Education and is currently doing a master's degree in Educational Measurement and Evaluation.
7	Man	She is a graduate in Elementary Mathematics Teacher Education and is currently doing a master's degree in Mathematics Teaching.
8	Man	She is a graduate in Elementary Mathematics Teacher Education and is currently doing a master's degree in Educational Measurement and Evaluation.
9	Woman	She is a graduate in Elementary Mathematics Teacher Education and is currently doing a master's degree in Educational Measurement and Evaluation.
10	Woman	She is a graduate in Elementary Mathematics Teacher Education and is currently doing a master's degree in Educational Measurement and Evaluation.

Of the eight questions in the final form of the test six involved the relational level of SOLO taxonomy. The other two questions involved the extended abstract level. Thus, the responses to six of eight questions on the test vary from prestructural level to the relational level, while the responses to the other two questions ranged between the prestructural level and the extended abstract level. Since the questions on the unidimensional and multidimensional levels of SOLO taxonomy are intended to assess recall and memorization, the researchers decided that asking these questions in the multiple choice format would be more economical than the open-ended format. For this reason, questions for the unidimensional and multidimensional levels were excluded from the study. The extended abstract level requires making generalizations, hypothesizing, induction, deduction and combinational reasoning processes and responding to questions on this level is closely related to students' development of abstract thinking. Considering the characteristics of cognitive development, the researchers believed that abstract thinking begins to be developed by eight grade students, but is not yet completely mastered (Erden & Akman, 2011). For this reason, the questions on the extended abstract level were included on the test, but limited to two questions. Since the questions on the relational level were aimed to assess the behaviors such as building correlations, analyzing and explaining causes and results, the questions on this level were deemed to be more suitable for the cognitive development of eight grade students. Thus, the test included more questions from the relational level than the extended abstract level. Table 3 presents two questions that exemplify the relational and extended abstract levels on the mathematics achievement test.

Table 3. Sample Questions for Relational and Extended Abstract Levels of SOLO Taxonomy on Mathematics Achievement Test

 For the ABD and CBD triangles on the left; when ǀABǀ=5 cm, ǀADǀ=12 cm, ǀBCǀ=6 cm and ǀCDǀ=8 cm, what is the range of possible values for ǀBDǀ?	This question asks the students to solve the triangle inequality for both triangles and then, combine these inequalities to reach a consistent whole. Therefore, this question corresponds to the relational level of the SOLO taxonomy.
Ali and Ayşe are playing a game, and trying to make houses that are right next to each other using toothpicks with equal length. Here are the houses Ali and Ayşe made while playing the game. Accordingly,  a) Calculate the number of toothpicks required to make 9 houses. b) If 169 toothpicks are needed to make 42 houses, how many toothpicks do they need to have to make 43 houses? c) Create an algebraic expression for the correlation between the number of houses and the number of toothpicks. d) Ali and Ayşe want to make houses with different geometrical shapes instead of *pentagon-shaped* ones, and want them to be right next to each other again. Decide on a different geometrical shape to help them. For the houses that have the geometrical shape you decided, create an algebraic expression between the number of toothpicks and the number of houses.	*Option a* for this question is unistructural, and it is sufficient for the students to draw six more houses next to these three houses to answer this option correctly. *Option b* is multistructural. The students who fail to build an algebraic correlation between the numbers of houses and toothpicks, yet calculate the common difference of the pattern can answer this option correctly. *Option c* requires that students should build an algebraic correlation between the numbers of houses and toothpicks. However, option c is included in the relational level since the correlation to be built is limited to the information given in the question. *Option d* asks the students to build a correlation beyond the information given to them, which puts it on the extended abstract level. Due to the hierarchical structure of the SOLO taxonomy, the highest step that is calculated is accepted as the basis when deciding on the cognitive level of a question. Thus, this question is included on the extended abstract level. However, it is possible to do partial scoring by considering the options that the student managed to answer correctly.

Study Sample

The sample of the study included seven mathematics teachers serving as raters (three females and four males) who scored the students' responses to the open-ended questions. The raters participated in the study on a voluntary basis, and they were selected according to the principle of accessibility. Table 4 presents the demographics of the raters. As Table 4 shows, one rater is an M.A. student in mathematics teaching, and the other six raters are M.A. students in assessment and evaluation in education. The researchers believe that the characteristics of raters' education levels are not a disadvantage for the generalizability of study results, since the study was focused on the comparison of the rubrics based on standard and SOLO taxonomies, and the same raters gave scores using both types of rubrics.

Table 4. The Demographics of the Raters

Rater	Gender	Age	Length of Service as a Teacher	Educational Status
R1	Woman	22	-	She is a graduate in Elementary Mathematics Teacher Education and is currently doing a master's degree in Mathematics Education.
R2	Woman	22	7 months	She is a graduate in Elementary Mathematics Teacher Education and is currently doing a master's degree in Educational Measurement and Evaluation.
R3	Woman	23	7 months	She is a graduate in Elementary Mathematics Teacher Education and is currently doing a master's degree in Educational Measurement and Evaluation.
R4	Man	26	2 years	He is a graduate in Elementary Mathematics Teacher Education and is currently doing a master's degree in Educational Measurement and Evaluation.
R5	Man	25	2 years	He is a graduate in Elementary Mathematics Teacher Education and is currently doing a master's degree in Educational Measurement and Evaluation.
R6	Man	25	7 months	He is a graduate in Elementary Mathematics Teacher Education and is currently doing a master's degree in Measurement and Evaluation.
R7	Man	26	3 years	He is a graduate in Mathematics Majored in Computer Science and is currently doing a master's degree in Educational Measurement and Evaluation.

Data Collection Tools

The authors used standard rubrics as well as rubrics based on the SOLO taxonomy to rate students' responses to open ended mathematical questions. Both were developed as task-specific holistic rubrics. The authors used separate standard and SOLO-based rubrics to rate every question on the mathematical achievement test. To compare their rater effects, the authors used identical ratings to avoid any influence that might be created by the use of different ratings in rubrics. The authors used a four level rating system for both the standard and the SOLO based rubrics for the items 1 to 6 in the mathematical achievement test. For items seven and eight, the authors used a five level rating system for both rubrics. Appendix-1 contains an example of an open-ended question on the test as and its standard and SOLO-based rubrics.

Standard Rubrics

The authors created a separate rubric for each item on the achievement test, which made eight standard rubrics in total. The authors used a four level system for six questions. For the other two questions, which included sub-targets, the authors used a five level system. Then, the authors consulted the opinions of five experts about these rubrics. Table 5 presents the demographics of the experts consulted about the standard rubrics.

Table 5. The Demographics of the Experts Consulted About the Standard Rubrics

Expert	Gender	Educational Status
1	Man	He is an associate professor in educational measurement and evaluation
2	Man	He is an associate professor in classroom teaching.
3	Woman	She has M.A. and Ph.D. degrees in Curriculum and Instruction.
4	Man	He is a graduate in Elementary Mathematics Teacher Education. He also has an M.A. degree in Curriculum and Instruction and is a Ph.D. student in the same field.
5	Man	He is a graduate in High School Mathematics Teacher Education. He also has an M.A. degree in Mathematics Education and is a Ph.D. student in the same field.

The expert opinions showed that: *i)* the statements in the rubrics were clear and understandable, *ii)* the rating categories were well-defined, *iii)* the differences between rating categories were clear, *iv)* the rubrics could be used to rate student groups with any levels of success, *v)* the rating criteria reflected all sides of the characteristic the authors aimed to measure and did not include any rating criteria other than that characteristic. Thus, the experts deemed that the standard rubrics were ready for use and did not require revision.

Rubrics Based on the SOLO Taxonomy

The authors also created eight rubrics based on the SOLO taxonomy. Six questions in the mathematical achievement test involve the relational level. The authors used a four level rating in the rubrics created for these questions. Those levels were *prestructural* (0), *unistructural* (1), *multistructural* (2) and *relational* (3). The other two questions in the test involve the extended abstract level. The responses that students could give to these questions range from prestructural level to the extended abstract level. Accordingly, the authors used a five level rating system for the SOLO-based rubrics for these items. These levels were *prestructural* (0), *unistructural* (1), *multistructural* (2), *relational* (3) and *extended abstract* (4). The authors consulted the opinions of four experts after creating the SOLO-based rubrics. Table 6 shows the demographics of the experts consulted about the SOLO-based rubrics.

Table 6. The Demographics of the Experts Consulted About the SOLO-Based Rubrics

Expert	Gender	Educational Status
1	Man	He is an associate professor in educational measurement and evaluation.
2	Woman	She has M.A. and Ph.D. degrees in Curriculum and Instruction.
3	Man	He is a graduate in Elementary Mathematics Teacher Education. He also has an M.A. degree in Mathematics Education.
4	Man	He is a graduate in Elementary Mathematics Teacher Education. He is also an M.A. student in Mathematics Educaytion and works as a Mathematics teacher.

The expert opinions indicated that: *i)* the statements in the rubrics were clear and understandable, *ii)* the rating categories were consistent with the levels of the SOLO taxonomy, *iii)* the differences between rating categories were clear, *iv)* the rubrics could be used to rate student groups with any levels of success, *iv)* the rating criteria reflected all sides of the characteristic the authors aimed to measure and did not include any rating criteria other than that characteristic. Thus, the experts deemed that the SOLO-based rubrics were ready for use and did not require revision.

Procedure

The data were collected in the 2014 spring semester. In the first phase, the mathematical achievement test, which was created to produce the documents to be evaluated by the raters, was administered to the students in their classroom. The students were asked to write down the mathematical operations they did to solve the problem in a clear way. After the test was administered, the authors numbered the exam sheets. The authors made 14 copies of the exam sheets since each of the seven raters rated the mathematical achievement test twice, using the standard rubric and the rubric based on SOLO taxonomy. This produced the documents that were to be evaluated by the raters. The scoring based on the standard rubric and the rubric based on SOLO taxonomy was done with these documents, which were the data source of the study.

For assessments which cannot be done objectively, it is necessary to inform the raters before they give their rates about the dimensions of the performance to be rated and the categories of the rubrics that will be used to assess that performance (Kutlu, Dogan, & Karakaya, 2010). Considering this necessity, the researchers gave training to the raters about the use of standard rubrics after deriving the documents to be used for scoring, which also included sample scoring by raters using the standard rubric. The authors' training program thus included the introduction of the standard rubrics and examples of such ratings. The sample scorings were done on four questions which were included in the pilot version of the mathematical achievement test conducted with 13 students but were not included in the actual implementation. For these four questions, the authors used students' responses that represented higher, moderate and lower achievement levels. The raters scored the students' answers based on the standard rubrics that were created for use in the sample ratings during the training session. Once the sample ratings were complete, the authors provided feedbacks to the raters about their assessments and completed the rater training on standard rubrics. After the training, the raters assessed the exam sheets of 104 students during two days to two weeks.

Then, the raters assessed the exam sheets using SOLO-based rubrics. The authors taught the raters about using SOLO-based rubrics. This training included the SOLO taxonomy and the SOLO-based rubrics. The raters also did practice ratings using the SOLO-based rubrics. As done in the first rater training, the sample ratings were conducted on the four questions that were included in the pilot version of the mathematical achievement test that was administered to 13 students but were not included not in the actual implementation. For these four questions, the authors used students' responses that represented higher, moderate and lower achievement levels. The raters evaluated the students' answers using rubrics based on SOLO taxonomy and were created to be used for these questions in the training. Once the sample ratings were complete, the authors provided feedback to the raters about their ratings and completed the rater training on the rubrics based on SOLO taxonomy. The raters did this round of assessments during eight to 22 days. The steps followed in the collection of research data was also summarized in Figure 1.

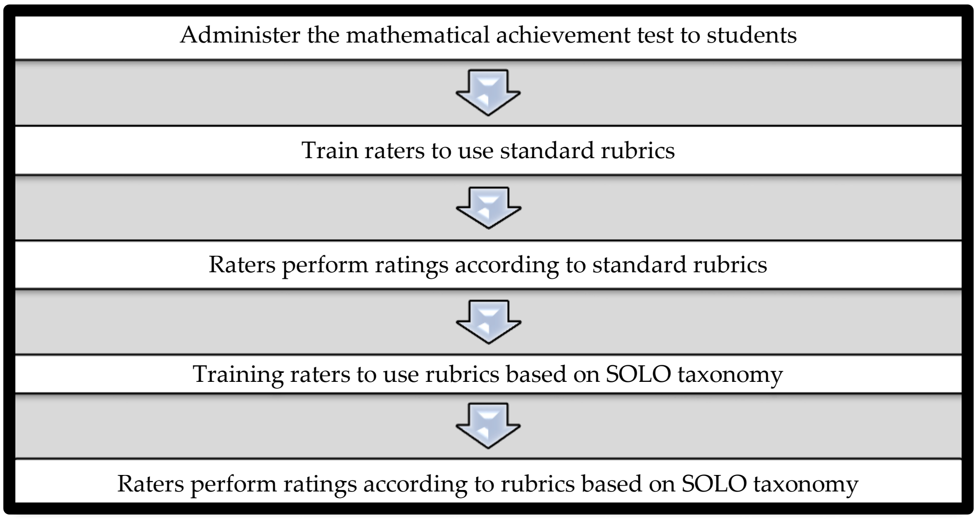

Administer the mathematical achievement test to students
Train raters to use standard rubrics
Raters perform ratings according to standard rubrics
Training raters to use rubrics based on SOLO taxonomy
Raters perform ratings according to rubrics based on SOLO taxonomy

Figure 1. The Procedures Followed During the Data Collection Process

Data Analysis

The study data were the ratings of 104 students' responses to eight open-ended mathematical questions by seven raters using standard and SOLO-based rubrics. Thus, there are three facets in the study. These facets are the students, the items and raters. The raters' assessments of the open-ended mathematical questions using standard and SOLO-based rubrics were analyzed using the many-facet Rasch model by means of FACETS program (Linacre, 2014). The study scored six questions on the mathematics achievement test using 4-point rubrics. The other two questions were scored using a 5-point rubric. Since the rating categories were different for the relational and extended abstract questions, the analyses used mixed rating scale forms. Before doing the many-facet Rasch analyses, the authors tested the correctness of the assumptions about these analyses. These assumptions include unidimensionality, local independence and the fit between model and data.

Unidimensionality

The authors did an exploratory factor analysis (EFA) to test whether the study data confirmed the unidimensionality assumption. The factor analysis was done considering the means of the raters' scores for each item. For the rating done using standard rubrics, it was found that there was a one-factor structure which explained 31.82% of the total variance, and the factor loads of the test items ranged between 0.40 and 0.74. For the ratings performed using the rubrics based on SOLO taxonomy, it was found that the test had a one-factor structure which explained 30.84% of the total variance, and the factor loads of the items ranged between 0.35 and 0.70. According to the EFA results, the unidimensionality assumption was confirmed.

Local Independence

Local independence is an assumption that works parallel to unidimensionality. Thus, when the unidimensionality assumption is confirmed, the local independence assumption is also confirmed (Hambleton, Swaminathan, & Rogers, 1991). Based on this point, the authors decided that the study findings confirmed the local independence assumption. In other words, the authors did not test the local independence assumption but accepted that it was confirmed since the unidimensionality assumption was also confirmed.

The Fit between Model and Data

The fit between the model and the data is determined by examining the standardized residual values (StRes) created by the many-facet Rasch analysis. According to Linacre (2014); in order for the model and the data to be fit with each other, the number of the StRes left out of the ±2 interval should not be more than 5% of the total data. Again according to Linacre (2014), in order for the model and the data to be fit with each other, the number of the StRes left out of the ±3 interval should not be more than 1% of the total data. In the study, 104 students' responses to eight questions were rated by seven raters. Thus, there were 5824 (104×8×7) data in total provided by the evaluations done using the standard rubrics and the rubrics based on the SOLO taxonomy. In the ratings done using the standard rubrics, the number of the StRes left out of ±2 interval was 271 (4.65%), and that of the StRes left out of ±3 interval was 56 (0.96%). Therefore, the fit between the model and the data in the rating done using standard rubrics was sufficient.

An analysis of the rating done based on the SOLO taxonomy indicated that the number of the StRes left out of ±2 interval was 289 (4.96%) and that of the StRes left out of ±3 interval was 91 (1.56%). Accordingly, the percentage of the StRes left out of ±3 interval was above the 1% criterion offered by Linacre (2014). However, Linacre (2014) did not precisely define these criteria, which he suggested ought to be considered when making decisions about the fit between the model and the data. Rather, he expressed them as approximate values. When the percentage of the StRes which are determined to be left out of the ±3 interval in the many-facet Rasch analysis is tackled this way, the fit between the model and the data is acceptable. Accordingly, McNamara (1996) said that the many-facet Rasch model should be used as long as the percentage of the StRes left out of ±2 or ±3 interval does not remarkably deviate from the values that are suggested to be accepted as the criteria. In basic item response theory, the analyses should be conducted with the model that better fits with the data set, whether one, two or three-parameter models. This means that the two-parameter model can be used when the three-parameter model is not fit enough with the data, or the one-parameter model can be used when the two-parameter model poorly fits with the study data. However, there is no alternative model that can be used instead when the fit between the model and the data is not high enough in the many-facet Rasch model. With respect to this point, it is suggested that the many-facet Rasch model should be used in performance evaluation even if the fit between the model and the data is not high enough (McNamara, 1996). Therefore, the percentage of the StRes left out of the ±3 interval in the ratings performed using rubrics based on the SOLO taxonomy is small enough to allow use of the many-facet Rasch model. After determining that the assumptions were confirmed, the authors performed the many-facet Rasch analysis. Following these analyses, the outputs were examined according to the criteria for rater severity and leniency in the literature. Statistical indicators at the group and individual levels which were analyzed to identify rater severity and leniency (Myford & Wolfe, 2004) were given on Table 7.

Table 7. Statistical Indicators of Rater Severity and Leniency

Group Level	**Individual Level**
- Separation ratio and reliability index in rater facet	- Any rater's location on the variable map differing from the other raters
- Statistically significant chi-square value in rater facet	- The *t* values calculated by using the logit measures for the raters, the mean and standard error of these measures being statistically significant for any one of the raters

As Table 7 shows, the significance of the chi square test result for the rater facet is the first indicator of the group-level rater severity and leniency. The significant chi square value shows that at least one rater scored more leniently or severely than the others. The other group-level indicators of rater severity and leniency are separation ratio and reliability index. The reliability index is expressed between 0 and 1, while the separation ratio is expressed between the interval of 1 and infinity. Although these two statistical values are reported as different metrics, they are both calculated using the same information and lead to similar results for a certain facet. Considering the item and person facets, the reliability index is interpreted similarly to the Cronbach's alpha internal consistency coefficient (Bond & Fox, 2007). Thus, it is suggested that the criteria of the reliability index should be at least 0.70, as in Cronbach's alpha coefficient (Walker, Engelhard, & Thompson, 2012). Values higher than 0.70 indicate that students with different ability levels can be distinguished successfully and that items on the assessment tool can be scored independently. The high values for the separation ratio and reliability index in the rater facet indicate that the consistency between the raters (inter-rater reliability) is weak and that differentiation is high. For this reason, it is desirable for the separation ratio and reliability index of the rater facet to be low. However, there is no clear criterion in the relevant literature for determining the maximum value the separation ratio and reliability index should have in order to decide that the raters have similar severity and leniency.

Although the group-level statistical indicators show any difference between raters' severity and leniency, they do not give any information about which rater or raters cause this difference, if any. It is necessary to examine the individual-level statistical indicators of rater severity and leniency to identify the rater that causes the difference. First of these individual-level indicators of rater severity and leniency is the *t* value which is calculated using the logit values of the each raters, the mean and standard error of these logit values. As seen in Table 7, another individual-level indicator of rater severity and leniency is raters' positions on the variable map. In the many facet Rasch analysis, all variability sources are converted to the logit scale with equal intervals, and presented together on the variable map. The accumulation of the raters in close positions on this line shows that they gave scores with similar severity and leniency. When the raters have different positions on the variable map, this means that they have different severity and leniency.

Results

This section will present the study's results. The outputs of many facet Rasch model contains numerous tables and figures related to person, item and rater facets. However, in the presentation of the results only the tables and figures that may be statistical indicators of rater severity and leniency will be given. First, the authors will describe the outcomes of the standard rubric ratings. Figure 2 shows the variable map for them.

```
+-----------------------------------------------------------------------+
|Measr| + EXAMİNEE                |-ITEM |+EXAMİNEE|-RATER     | S.1 | S.2 |
|-----+-------------------------+------+--------+-----------+-----+-----|
|   1 +                         +      +        +           + (3) + (4) |
|     | 87                      |      | *      |           |     |     |
|     |                         | 1    |        |           |     |     |
|     | 37                      |      | *      |           |     |     |
|     |                         | 6    |        |           |     |     |
|     | 98                      |      | *      |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     | 79  93  97  100         | 3    | ****   |           |     |     |
|     |                         |      |        |           |     | --- |
|     | 24  27                  |      | **     |           |  2  |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        | 6         |     |     |
|     |                         |      |        |           |     |     |
|     | 4   28  67              | 4    | ***    |           |     |     |
|     | 2   9   94  104         |      | ****   | 4         |     |     |
|     | 6   11  13  73          |      | ****   | 7         |     |     |
|     |                         |      |        |           |     |     |
|     | 5   83                  |      | **     |           |     |     |
|     | 8   30                  |      | **     |           |     |     |
*   0 * 1   89                  *      * **     * 5         *     *     *
|     | 10  12  46  57  58  96  |      | ****** |           | --- |     |
|     | 55  64  69  74  95      |      | *****  |           |     |  2  |
|     | 38  86                  |      | **     | 2         |     |     |
|     | 35  71  77              |      | ***    |           |     |     |
|     | 44  75  85  91          |      | ****   | 1         |     |     |
|     | 3   29  36  70          |      | ****   |           |     |     |
|     | 49  53  56  76  81      | 7  8 | *****  |           |     |     |
|     | 21  22  63              |      | ***    |           |     |     |
|     | 7   42  54              |      | ***    | 3         |     |     |
|     | 14  61  62              |      | ***    |           |     |     |
|     | 31  40  47  48  66      |      | *****  |           |     |     |
|     | 90  103                 |      | **     |           |  1  |     |
|     | 34                      |      | *      |           |     |     |
|     | 16  84                  |      | **     |           |     |     |
|     | 18  33  41              | 2    | ***    |           |     | --- |
|     | 39  80  99  101         |      | ****   |           |     |     |
|     | 19                      |      | *      |           |     |     |
|     | 45  59  72              |      | ***    |           |     |     |
|     | 32  52                  |      | **     |           |     |     |
|  -1 + 51                      +      + *      +           +     +     |
|     | 88                      |      | *      |           |     |     |
|     |                         | 5    |        |           |     |     |
|     | 50  68  78  102         |      | ****   |           |     |     |
|     |                         |      |        |           |     |     |
|     | 65                      |      | *      |           |     |     |
|     |                         |      |        |           | --- |  1  |
|     |                         |      |        |           |     |     |
|     | 25                      |      | *      |           |     |     |
|     | 20                      |      | *      |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     | 82                      |      | *      |           |     |     |
|     | 17                      |      | *      |           |     |     |
|     |                         |      |        |           |     |     |
|     | 23  43                  |      | **     |           |     |     |
|     | 26                      |      | *      |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|  -2 + 15                      +      + *      +           +     + --- |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     | 60                      |      | *      |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     | 92                      |      | *      |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|     |                         |      |        |           |     |     |
|  -3 +                         +      +        +           + (0) + (0) |
|-----+-------------------------+------+--------+-----------+-----+-----|
|Measr|+EXAMİNEE                |-ITEM | * = 1  |-RATER     | S.1 | S.2 |
+-----------------------------------------------------------------------+
```

Figure 2. The Rasch Model Variable Map of the Ratings Done with Standard Rubrics

The measurements of the raters are shown in the fifth column of Figure 2. The raters at the top of the column with a high logit score are more severe, while those at the bottom of the column with a low logit score are more lenient. Rater number 6 was the most severe rater with 0.39 logit, and rater number 3 was the most lenient rater with -0.45 logit. There are differences of severity and leniency between the raters, which is indicated by the measurements range from the negative to the positive end of the logit scale in the rater facet. The measurement reports of the rater facet should be examined to make a definitive judgment. Table 8 presents the measurement reports of the rater facet.

Table 8. The Measurement Reports for Rater Facet of the Ratings Done with Standard Rubrics

Rater	Measure	Model S.E.	Infit MnSq	Outfit MnSq
R6	.39	.04	.98	1.08
R4	.26	.04	1.05	1.09
R7	.19	.04	.76	.90
R5	.02	.04	.97	1.01
R2	-.16	.04	.87	.97
R1	-.24	.04	1.16	1.21
R3	-.45	.04	1.12	1.12
Mean	.00	.04	.99	1.05
Standard Deviation (Population)	.28	.00	.13	.10
Standard Deviation (Sample)	.30	.00	.14	.10
Model, Population: RMSE=.04	Standard Deviation =.27	Separation =6.53	Reliability =.98	
Model, Sample: RMSE=.04	Standard Deviation =.30	Separation =7.07	Reliability =.98	
Model, Fixed (all same) chi-square =306.0	df=6 p=.00			
Model, Random (normal) chi-square =5.9	df=5 p=.32			

Table 8 shows that the logit measurements of the raters range between 0.39 and -0.45, and that the interval of rater severity and leniency is 0.84 logits [.39-(-.45)]. The logit value reported for each rater and raters' intervals regarding the logit criteria are also shown on the variable map. Another statistics in Table 8 is infit and outfit mean square statistical values. When the mean of the infit and outfit mean square is 1, it shows that the fit between the data and the model is flawless. However, it is usually impossible that the fit between the model and the data will be flawless in actual measurements (Brentari & Golia, 2008). Therefore, the acceptable interval of the infit and outfit mean square statistics should be determined. Wright and Linacre (1994) reported that the infit and outfit mean square values between 0.6 and 1.4 are acceptable. By this criterion, the values below 0.5 and those above 1.5 indicate that the data are not suitable for measurement. However, Myford and Wolfe (2003) say that adequacy values up to 2 are all acceptable. According to Myford and Wolfe (2003), infit and outfit mean square statistics below 2 are acceptable, while values between 1.5 and 2 are not useful for measurement, but are not harmful. Infit and outfit statistics above 2 indicate that the data are unfavorable for the measurement (Sudweeks, Reeve, & Bradshaw, 2004). The means of the infit and outfit mean square statistical values reported for the raters are 0.99 and 1.05 which are very close to 1. These values show that the data are consistent with the model. The authors also found that the infit and outfit mean square statistics were within the acceptable interval for all raters, so no raters had a negative influence on the fit between the model and the data.

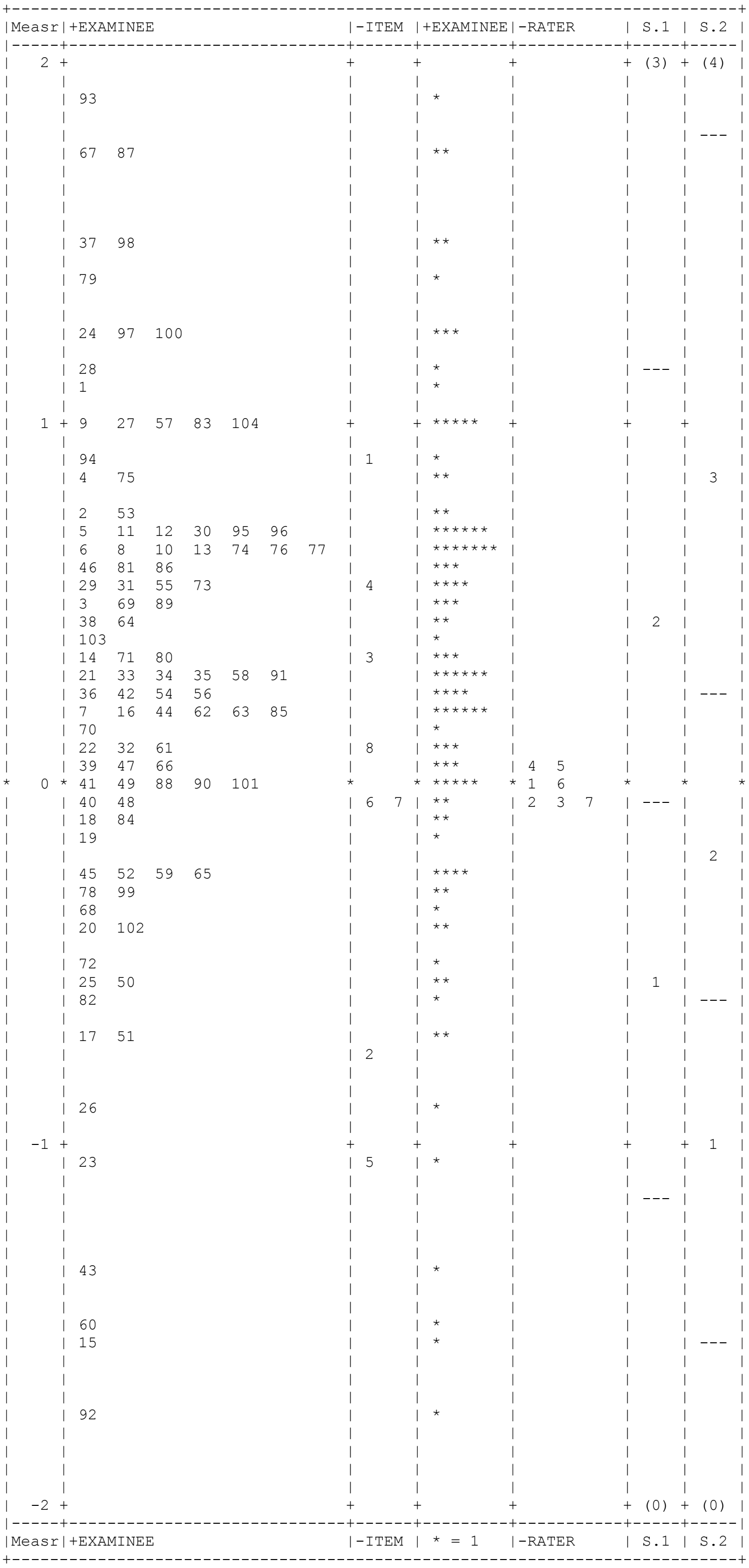
A review of the separation ratio and reliability index in the rater facet shows that there are two different models, which are the population and the sample. According to Linacre (2014), the separation ratio and reliability index in the "*model, population*" line should be considered if all possible components of any facet is included in the model. For instance, if the gender variable is a facet included in the analysis, then all possible components of this facet will be included in the model as "male/female." In such a case, the authors consider the separation ratio and reliability index in the "*model, population*" line. However, the separation ratio and reliability index in the "*model, sample*" line is considered only if a randomly chosen part of all components of the facet are included in the model. For instance, it is not possible to include all possible components of person, rater, or item facets in the model. For these facets, the components randomly chosen from person, rater and item population are included in the model. In such a case, it is required to interpret the separation ratio and reliability index in the "*model, sample*" line (Linacre, 2014). Accordingly, the separation ratio and reliability index in the "*model, sample*" line were considered when interpreting the findings in the rater facet. Table 8 shows that the separation ratio of the rater facet is 7.07 and the reliability index is 0.98. The separation ratio and reliability index are statistics that are related to the reliability of the difference. The high reliability value for the person facet shows that the students with different ability levels can be distinguished effectively. The high reliability index for the item facet shows that the different conceptual aspects of the characteristic to be assessed can be distinguished by raters. High reliability value for the rater facet shows that the raters are different from each other regarding their severity and leniency since the reliability index calculated for the rater facet shows the difference between the raters rather than similarity between them (Haiyang, 2010). Accordingly, it is favorable that the distinction rate and reliability index in the rater facet are low, which is in contrast with the item and person facets. A reliability of 0.98 shows that the raters do differ. The chi-square value determines whether this difference is significant or not. In the many-facet Rasch analysis, there are two different chi-square values reported, which are *random normal* and *all the same*. The *random normal* chi-square value is accepted as a reference to decide if the components of any facet represent a sample randomly chosen from a population with a normal distribution. The *all the same* chi-square value is examined to determine any significant difference between the components of the facet after the measurement error is allowed (Linacre, 2014). Accordingly, the authors analyzed the all the same chi-square value to see any significant difference between the raters by leniency and severity. Since the chi-square value was statistically significant [χ^2=306.00, *sd*=6, *p*<.01], there is a significant difference between raters in terms of severity and leniency.

Given this difference, the researchers needed to determine which rater or raters caused this difference by doing individual analyses. The variable map is a statistical indicator of rater severity and leniency on the individual level. The source of the difference between raters is indicated by outlying locations in the rater column of the variable map (Myford & Wolfe, 2004). The variable map on Figure 2 shows that all raters have different locations on the logit scale. Rater 5 is at the 0 level of the logit scale, while raters 4, 6 and 7 are on its positive end, and raters 1, 2 and 3 are on its negative end. This shows that Raters 4, 6 and 7 are more severe, while Raters 1, 2 and 3 are more lenient. However, the *t*-values for each rater should be computed to make a definitive judgment on this issue. To compute the *t*-value, the logit mean score of all raters is subtracted from the logit measurement of any rater and the result is divided by the standard error of logit measurements. Then, the significance test is performed by comparing this *t*-value to the critical *t*-value of the relevant degree of freedom. Since there are seven raters included in the study, the degree of freedom was 7-1=6, and the critical *t*-value at the 0.01 level with this degree of freedom was found to be 3.71.

Table 9. The Results of the *t* test on the Significance of the Differences in Severity and Leniency between Raters Scoring with Standard Rubrics

Rater	***t*-values**	**The significance of the difference**
R6	9.75	$\lvert t_{calculated} \rvert > t_{critical}$; therefore, the difference is significant. Raters 4, 6 and 7 are located on the positive side of the variable map. This means, it was found that these raters were significantly more severe than the other raters.
R4	6.50	
R7	4.75	
R5	.05	$\lvert t_{calculated} \rvert < t_{critical}$; therefore, the difference is not significant.
R2	-4.00	$\lvert t_{calculated} \rvert > t_{critical}$; therefore, the difference is significant. Raters 1, 2 and 3 are located on the negative side of the variable map. Thus, it was found that these raters were significantly more lenient than the other raters.
R1	-6.00	
R3	-11.25	

Table 9 shows the *t* values for each rater in the study and the results about the significance of these values. According to this table, raters 4, 6 and 7 are more severe, while raters 1, 2 and 3 are more lenient. The authors also analyzed the SOLO-based ratings. Figure 3 shows the variable map for them.



```
+----------------------------------------------------------------------------+
|Measr|+EXAMINEE                     |-ITEM |+EXAMINEE|-RATER     | S.1 | S.2 |
|-----+-----------------------------+------+---------+-----------+-----+-----|
|   2 +                             +      +         +           + (3) + (4) |
|     |                             |      |         |           |     |     |
|     | 93                          |      | *       |           |     |     |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           |     | --- |
|     | 67  87                      |      | **      |           |     |     |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           |     |     |
|     | 37  98                      |      | **      |           |     |     |
|     |                             |      |         |           |     |     |
|     | 79                          |      | *       |           |     |     |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           |     |     |
|     | 24  97  100                 |      | ***     |           |     |     |
|     |                             |      |         |           |     |     |
|     | 28                          |      | *       |           | --- |     |
|     | 1                           |      | *       |           |     |     |
|     |                             |      |         |           |     |     |
|   1 + 9   27  57  83  104         +      + *****   +           +     +     |
|     |                             |      |         |           |     |     |
|     | 94                          | 1    | *       |           |     |     |
|     | 4   75                      |      | **      |           |     |  3  |
|     |                             |      |         |           |     |     |
|     | 2   53                      |      | **      |           |     |     |
|     | 5   11  12  30  95  96      |      | ******  |           |     |     |
|     | 6   8   10  13  74  76  77  |      | ******* |           |     |     |
|     | 46  81  86                  |      | ***     |           |     |     |
|     | 29  31  55  73              | 4    | ****    |           |     |     |
|     | 3   69  89                  |      | ***     |           |     |     |
|     | 38  64                      |      | **      |           |  2  |     |
|     | 103                         |      | *       |           |     |     |
|     | 14  71  80                  | 3    | ***     |           |     |     |
|     | 21  33  34  35  58  91      |      | ******  |           |     |     |
|     | 36  42  54  56              |      | ****    |           |     | --- |
|     | 7   16  44  62  63  85      |      | ******  |           |     |     |
|     | 70                          |      | *       |           |     |     |
|     | 22  32  61                  | 8    | ***     |           |     |     |
|     | 39  47  66                  |      | ***     | 4  5      |     |     |
*   0 * 41  49  88  90  101         *      * *****   * 1  6      *     *     *
|     | 40  48                      | 6  7 | **      | 2  3  7   | --- |     |
|     | 18  84                      |      | **      |           |     |     |
|     | 19                          |      | *       |           |     |     |
|     |                             |      |         |           |     |  2  |
|     | 45  52  59  65              |      | ****    |           |     |     |
|     | 78  99                      |      | **      |           |     |     |
|     | 68                          |      | *       |           |     |     |
|     | 20  102                     |      | **      |           |     |     |
|     |                             |      |         |           |     |     |
|     | 72                          |      | *       |           |     |     |
|     | 25  50                      |      | **      |           |  1  |     |
|     | 82                          |      | *       |           |     | --- |
|     |                             |      |         |           |     |     |
|     | 17  51                      |      | **      |           |     |     |
|     |                             | 2    |         |           |     |     |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           |     |     |
|     | 26                          |      | *       |           |     |     |
|     |                             |      |         |           |     |     |
|  -1 +                             +      +         +           +     +  1  |
|     | 23                          | 5    | *       |           |     |     |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           | --- |     |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           |     |     |
|     | 43                          |      | *       |           |     |     |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           |     |     |
|     | 60                          |      | *       |           |     |     |
|     | 15                          |      | *       |           |     | --- |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           |     |     |
|     | 92                          |      | *       |           |     |     |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           |     |     |
|     |                             |      |         |           |     |     |
|  -2 +                             +      +         +           + (0) + (0) |
|-----+-----------------------------+------+---------+-----------+-----+-----|
|Measr|+EXAMINEE                     |-ITEM | * = 1   |-RATER     | S.1 | S.2 |
+----------------------------------------------------------------------------+
```

Figure 3. The Rasch Model Variable Map of the Ratings Done with Rubrics Based on the SOLO Taxonomy

Figure 3 shows that most raters are located at the 0 level of the logit scale or very close it. Since the raters are all in the central part of the logit scale, there is no remarkable difference between the raters in terms of severity and leniency. However, it is necessary to examine the measurement reports of the rater facet to determine any significant difference between the raters. Table 10 shows the measurement reports of the rater facet.

Table 10. The Measurement Reports for Rater Facet of the Ratings Done with SOLO-Based Rubrics

Rater	Measure	Model S.E.	Infit MnSq	Outfit MnSq
R6	.06	.04	1.02	1.19
R4	.04	.04	1.04	1.16
R7	.02	.04	.98	1.12
R5	.02	.04	1.00	1.15
R2	-.04	.04	1.04	1.12
R1	-.05	.04	.96	1.06
R3	-.06	.04	.91	1.00
Mean	.00	.04	.99	1.12
Standard Deviation (Population)	.04	.00	.04	.06
Standard Deviation (Sample)	.05	.00	.05	.07
Model, Population: RMSE=.04	Standard Deviation =.02		Separation =.51	Reliability =.21
Model, Sample: RMSE=.04	Standard Deviation =.03		Separation =.69	Reliability =.32
Model, Fixed (all same) chi-square =8.8 df=6 p=.18				
Model, Random (normal) chi-square =3.6 df=5 p=.61				

According to Table 10, the logit measurements of the raters range between 0.06 and -0.06 and the interval of raters' severity and leniency is 0.12 logits [.06-(-.06)]. This small interval, shows that the severity and leniency differences between the raters are slight. The means of the infit and outfit mean square statistical values reported for the raters are 0.99 and 1.12 which are very close to 1. These values show that the data are consistent with the model. Moreover, the infit and outfit mean square statistics of all raters are in the acceptable interval between 0.5 and 2.00 (Myford & Wolfe, 2003). This indicates that none of the raters had a negative influence on the fit between the model and the data.

Table 10 shows that the separation ratio of the rater facet is 0.69 and the reliability index of this facet is 0.32. The separation ratio and the reliability index in the rater facet are low, both of which indicate that there is no difference between the raters in terms of severity and leniency. However, the final decision on this issue is made using the chi-square value, which reflects any statistical significance of the difference between the raters (Linacre, 2014). Since the chi-square value is not statistically significant [χ^2=8.8, *sd*= 6, *p*>.05], it can concluded that there is no significant difference between the raters' severity or leniency. After the evaluation of the group statistics, the authors analyzed the statistical indicators on the individual level. Raters' locations on the variable map are the statistical indicators of rater severity and leniency on the individual level. The variable map on Figure 3 shows that the raters are all located at the 0 level of the logit scale, indicating highly consistent ratings. Another statistical indicator of the rater severity and leniency on the individual level is the *t*-values calculated for the participating raters. The *t*-values were calculated using the raters' logit measurements, the mean and the standard error of these measurements.

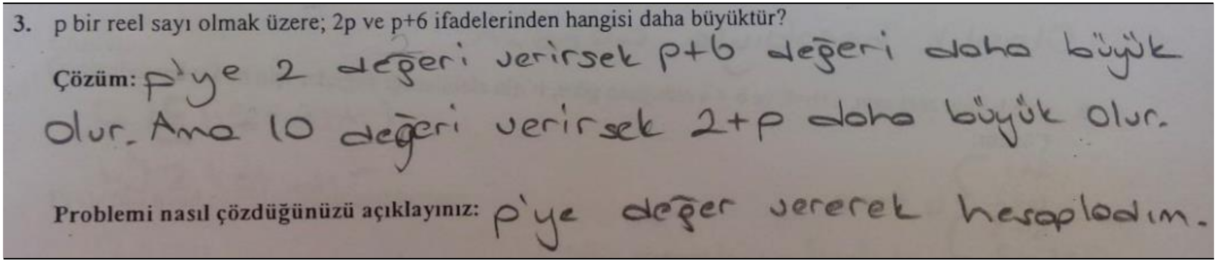
Table 11. The Results of the *t* test on the Significance of the Differences in Severity and Leniency between Raters Scoring with SOLO-Based Rubrics

Rater	*t*-values	The significance of the difference
R5	1.50	$\lvert t_{calculated} \rvert < t_{critical}$; therefore, the difference is not significant.
R4	1.00	
R1	.5	
R6	.5	
R3	-1.00	
R7	-1.25	
R2	-1.50	

Table 11 shows that the *t*-values range between -1.50 and 1.50. Since there are seven raters included in the study, the degree of freedom was 7-1=6, and the critical *t*-value at the 0.01 level with this degree of freedom was found to be 3.71. The calculated *t*-values are not above the critical *t*-value, which shows that there is no difference between the raters in terms of severity and leniency.

The sample answer presented in Table 12 shows that there is significant difference between raters when they give their scores using standard rubrics, but have similar severity and leniency when they give scores using the rubrics based on the SOLO taxonomy. Table 12 presents one of the student's responses to question 3 on the mathematics achievement test (see Attachment 1) together with the scores assigned to this response by the seven raters included in the question.

Table 12. A Sample Scoring by the Raters Based on the Standard Rubric, and on the Rubric Based on SOLO Taxonomy



3. p bir reel sayı olmak üzere; 2p ve p+6 ifadelerinden hangisi daha büyüktür?

Çözüm: p'ye 2 değeri verirsek p+6 değeri daha büyük olur. Ama 10 değeri verirsek 2+p daha büyük olur.

Problemi nasıl çözdüğünüzü açıklayınız: p'ye değer vererek hesapladım.

	R1	R2	R3	R4	R5	R6	R7
Standard Rubric	2	0	1	1	1	0	1
SOLO Based Rubric	2	2	2	2	2	2	2

Table 12 shows that the raters gave quite different scores to the same response when they used the standard rubric. The same response was assigned to the *good (2) category* of the standard rubric by rater 1, to *the acceptable (1) category* by raters 3, 4 and 5 and to the *inadequate (0) category* by rater 6. Using rubrics based on the SOLO taxonomy, the raters agreed that this response was included in the multistructural level of the SOLO taxonomy, and they all assigned it the same score.

Discussion

This study used the many-facet Rasch model to analyze rater severity and leniency for open-ended mathematical questions rated through standard and SOLO-based rubrics. Its results indicated that the agreement of the raters was low when they used standard rubrics and there was significant difference between raters by their severity and leniency. One of the main intentions of rubric use in performance-based assessments is to ensure that the ratings do not vary by rater (Moskal & Leydens, 2000; Purpura, 2004). In other words, rubrics should minimize the rater effect and increase inter-rater agreement (Dunbar, Brooks, & Miller, 2006). However, this study's results show that standard rubrics do not meet these expectations sufficiently. This finding is supported the results of the study conducted by Güler and Gelbal (2010). In their study, students' responses to open-ended mathematical questions were rated by four different raters using holistic rubrics created without using taxonomy. Their results showed that the agreement between raters was low, and that there was significant difference between raters' severity and leniency. Accordingly, this study's findings are consistent with those of the study by Güler and Gelbal (2010). However, the interpretation of this consistency should take the differences between the two studies into consideration. First of all, the study by Güler and Gelbal (2010) created a *general rubric* and used it to score all of the items, while this study created a specific rubric for each question in the mathematics achievement test, and used these *task-specific* rubrics to score the responses to the test questions. In addition, Güler and Gelbal used rubrics with six level rating. The standard rubrics used in this study were, though, had four and five level rating. Thus, neither task-specific nor general, neither six nor five or four level rating standard rubrics are completely effective at eliminating the differences in severity and leniency between raters. On the other hand, considering that both this study and the study conducted by Güler and Gelbal (2010) used standard rubrics with holistic structure, the argument about the effect of standard rubrics on rater severity and leniency may not be valid for analytical rubrics.

The individual-level indicators of rater severity and leniency demonstrated that three out of the seven raters in the study were lenient scorers, while three of them were severe. In the scores given based on standard rubrics, rater severity and leniency were observed in almost all of the raters in the study, which is consistent with the theoretical knowledge in the relevant literature. Accordingly, Cronbach (1990) said that rater severity and leniency was the most important rater effect in the rating process.

The authors also used the many-facet Rasch model to analyze ratings of students' responses to open-ended mathematical questions with SOLO-based rubrics. This analysis showed that the consistency of the raters was high and found no significant difference between raters' severity and leniency. This result showed that the SOLO-based rubrics helped eliminate differences between the raters and contributed to the objectivity of the rating process. The theoretical knowledge in the relevant literature claims that SOLO taxonomy makes rating measurements more clear (Hattie & Purdie, 1998) and consists of levels which can be understood easily (Biggs & Collis, 1982). The present study offers empirical evidences for this kind of theoretical knowledge.

An analysis of the empirical studies in the relevant literature indicates that some studies support the idea that SOLO-based rubrics increase the reliability of raters, while others conflict with this finding. For instance, the studies conducted by Burnett (1999) and Hundzynski (2008) analyzed the rater reliability of assessments done using the SOLO taxonomy. The reliability coefficients of the raters were 0.85 and 0.87, respectively, and in accordance with this study's findings. Another study that supports the results of this study was conducted by Yazıcı (2013). In the study conducted by Yazıcı (2013), three raters rated open-ended physics questions using rubrics based on the SOLO taxonomy. Its findings showed that the level of the reliability between raters was high, and that SOLO-based rubrics reduced the differences between the raters. Thus, this study's findings are similar to those of the studies mentioned. This is, though, not a complete overlapping, since the difference between raters was tested using many-facet Rasch model in this study, while the previous studies of this subject used correlation analysis for the same objective. Correlation analysis reveals whether the ranking by raters of the persons they evaluate is consistent (relative consistency), yet it does not give any information about the absolute consistency between them (Goodwin, 2001). On the other hand, in the many facet Rasch model the difference between raters regarding severity and leniency is calculated considering the true values of their scores for the persons they evaluate (absolute consistency) rather than their ranking of these persons (Sudweeks et al., 2004). Accordingly, a comprehensive analysis of this study's findings and those of previous studies of this subject shows that the rubrics based on SOLO taxonomy improve both relative and absolute consistency between raters.

The results of the study by Leung (2000) differ. Leung (2000) evaluated rater reliability of ratings done with SOLO-based rubrics and found that the correlation coefficient between the raters was 0.49. According to Leung (2000), their reliability was low because the raters were not used to rating with SOLO-based rubrics. The authors believe that the difference between this study and Leung's (2000) can be explained by the fact that the raters in this study were taught how to use SOLO-based rubrics. Presumably, the sample ratings done during the training helped them familiarize themselves with the rubric categories.

Suggestions

The results of this study suggest that rubrics based on the SOLO taxonomy should be used for the rating of open-ended mathematical questions both in large scale examinations and in classroom evaluations. The use of SOLO-based rubrics for rating open-ended mathematical questions will minimize the variance resulting from the rater. Moreover, the SOLO taxonomy can be used to rate open-ended questions in a variety of disciplines since it does not depend on the content (Kanuka, 2011). These results confirm that the study findings will make great contributions to the relevant practical studies.

This research puts forward both practical suggestions and implications for future research. There are no studies in the relevant literature identifying either standard rubrics or the rubrics based on taxonomies such as SOLO, Bloom, Fink, Detmer or Haladyana as more influential in reducing differences between raters. The authors believe that this study will meet this need in the literature since it does a comparative analysis of rater severity and leniency with standard and SOLO-based rubrics. However, this study does not explore whether rubrics based on the Bloom, Fink, Detmer and Haladyana taxonomies are better than standard rubrics at reducing rater effect. In this context, the assessment of the open-ended mathematical questions rated with standard rubrics and with any of these four taxonomies should be compared for rater effect. This study used the original five-level structure suggested by Biggs and Collis (1982) to prepare the SOLO-based rubrics it used. There are also studies which restructure the SOLO taxonomy by adding seventh, eighth, and ninth levels (Burnett, 1999; Chan et al., 2002). These studies show that the number of levels used in SOLO-based rubrics influences the assessment results. Therefore, future studies should analyze rater severity and leniency in open-ended mathematical questions assessed with SOLO-based rubrics that have more levels. Third, the rater effects analyzed in this study are limited to rater severity and leniency. Future studies should also analyze other rater effects such as central tendency, halo effect, rater bias and randomness. Another limitation is that the study consisted of 104 students' responses to eight open-ended mathematical questions rated by seven raters. In Rasch analyses, the findings provided by 100 to 200 students are accepted to be sufficient for parameter predictions. However, considering that analyses based on item response theory produce more accurate prediction with more participants (DeMars, 2010) and that the many-facet Rasch model is a continuation of the item response theory, it is suggested that similar studies should be conducted using a larger data source. Finally, this study analyzed the ability of standard rubrics and SOLO-based rubrics to reduce rater effects in open-ended mathematical questions. Similar studies should be conducted in different courses to be able to generalize its findings.

References

Airasian, P. W. (2005). *Classroom assessment*. New York: McGraw-Hill.

Baird, J. A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). *Marker effects and examination reliability a comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling*. Retrieved from http://www.ofqual.gov.uk/files/2013-01-21-marker-effects-and-examination-reliability.pdf

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* (Unpublished doctoral dissertation). University of Toronto, Canada. Retrieved from http://search.proquest.com/docview/304360302/fulltextPDF/4AEA7C68D8F945FEPQ/1?accountid=15780

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. Academic Press.

Bingölbali, E., Özmantar, M. F., & Akkoç, H. (2008). *Sınıf öğretmenlerinin farklı matematiksel çözüm yollarını değerlendirme süreçleri*. Paper presented at VII. Ulusal Sınıf Öğretmenliği Sempozyumu, Çanakkale, Turkey. Retrieved from http://mimoza.marmara.edu.tr/~hakkoc/yayin2008_bingolbali_ozmantar_akkoc_usos.pdf

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Brabrand, C., & Dahl, B. (2009). Using the SOLO taxonomy to analyze competence progression of university science curricula. *Higher Education*, *58*(4), 531-549. doi:10.1007/s10734-009-9210-4

Brentari, E., & Golia, S. (2008). Measuring job satisfaction in the social services sector with the Rasch model. *Journal of Applied Measurement*, *9*(1), 45-56. Retrieved from http://www.unibs.it/sites/default/files/ricerca/allegati/10061.pdf

Burnett, P. C. (1999). Assessing the structure of learning outcomes from counselling using the SOLO taxonomy: An exploratory study. *British Journal of Guidance & Counselling*, *27*(4), 567-580. doi:10.1080/03069889908256291

Chan, C. C., Hong, J. H., & Chan, M. Y. C. (2001). *Applying the structure of the observed learning outcomes (SOLO) taxonomy on student's learning outcomes: A comparative review*. Unpublished manuscript, Hong Kong Polytechnic University, Hong Kong.

Chan, C. C., Tsui, M. S., Mandy, Y. C., & Hong, J. H. (2002). Applying the structure of the observed learning outcomes (SOLO) taxonomy on student's learning outcomes: An empirical study. *Assessment and Evaluation in Higher Education*, *27*(6), 511-527. doi:10.1080/0260293022000020282

Collis, K. F., & Romberg, T. A. (1992). *Collis-Romberg mathematical problem solving profiles*. Melbourne: Australian Council for Educational Research.

Cronbach, L. I. (1990). *Essentials of psychological testing*. New York: Harper and Row.

Çetin, B., Boran, A., & Yazıcı, N. (2014). Fizik eğitiminde başarının ölçülmesinde SOLO taksonomisine göre hazırlanan rubriklerin incelenmesi. *Bayburt Üniversitesi Eğitim Fakültesi Dergisi*, *9*(2), 32-71. Retrieved from http://edergi.bayburt.edu.tr/index.php/befd/article/view/9/6

DeMars, C. (2010). *Item response theory*. Oxford, UK: Oxford University Press.

Dunbar, N. E., Brooks, C. F., & Miller, T. K. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education*, *31*(2), 2006, 115-128. doi:10.1007/s10755-006-9012-x

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197-221. doi:10.1207/s15434311laq0203_2

Erden, M., & Akman, Y. (2011). *Eğitim psikolojisi*. Ankara: Arkadaş Yayınevi.

Farrokhi, F., Esfandiari, R., & Vaez Dalili, M. (2011). Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal*, *15*, 70-77. Retrieved from http://www.idosi.org/wasj/wasj15(IPLL)11/12.pdf

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science*, *5*(1), 13-34. doi:10.1207/S15327841MPEE0501_2

Gronlund, N. E. (1998). *Assessment of student achievement*. Boston: Allyn and Bacon.

Güler, N. (2008). *Klasik test kuramı, genellenebilirlik kuramı ve Rasch modeli üzerine bir araştırma* (Unpublished doctoral dissertation). Hacettepe University, Institute of Social Science, Ankara, Turkey.

Güler, N., & Gelbal, S. (2010). Klasik test kuramı ve çok değişkenlik kaynaklı Rasch modeli üzerine bir çalışma. *Eğitim Araştırmaları Dergisi*, *38*, 108-125. Retrieved from http://www.aniyayincilik.com.tr/main/pdfler/38/7_guler_nese.pdf

Haiyang, S. (2010). An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics*, *33*(2), 87-102. Retrieved from http://www.celea.org.cn/teic/90/10060807.pdf

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications, Inc.

Hattie, J. A., & Purdie, N. (1998). The SOLO method and item construction. In G. Boulton-Lewis & B. Dart (Eds.), *Learning in Higher Education*. Hawthorn, Australia: ACER.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it?. *Psychological Methods*, *5*(1), 64-86.

Hundzynski, C. (2008). *Elementary teachers in a science inquiry study group: Concerns, uses, and reflections* (Unpublished doctoral dissertation). Fordham University, New York, ABD. Retrieved from http://search.proquest.com/docview/304641444/previewPDF/AB40E91C649C453CPQ/1?accountid=15780

İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli Rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(2), 346-368. doi:10.16986/HUJE.2016015182

Jackson, S. E., Schuler, R. S., & Werner, S. (2009). *Managing human resources*. Mason, OH: Cengage/Southwestern Publishers.

Jurdak, M. (1991). Van Hiele levels and the SOLO taxonomy. *International Journal of Mathematical Education in Science and Technology*, *22*(1), 57-60. doi:10.1080/0020739910220109

Kan, A. (2005). Yazılı yoklamaların puanlanmasında puanlama cetveli ve yanıt anahtarı kullanımının (farklı) puanlayıcı güvenirliğine etkisi. *Eğitim Araştırmaları Dergisi*, *19*, 207-219. Retrieved from http://www.ejer.com.tr/0DOWNLOAD/pdfler/tr/821760610.pdf

Kan, A. (2007). Performans değerlendirme sürecine katkıları açısından yeni program anlayışı içerisinde kullanılabilecek bir değerlendirme yaklaşımı: Rubrik puanlama yönergeleri. *Kuram ve Uygulamada Eğitim Bilimleri*, *7*(1), 129-152. Retrieved from https://www.edam.com.tr/kuyeb/pdf/en/567aeeee08a7e62db0b82fd5312c9d7baneng.pdf

Kanuka, H. (2011). Interaction and the online distance classroom: Do instructional methods effect the quality of interaction?. *Journal of Computing in Higher Education*, *23*(2-3), 143-156. doi:10.1007/s12528-011-9049-4

Kind, P. M. (1999). Performance assessment in science-What are we measuring?. *Studies in Educational Evaluation*, *25*(3), 179-194. doi:10.1016/S0191-491X(99)00021-8

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1992). *The reliability of scores from the 1992 Vermont portfolio assessment program* (Center fort he Study the Evaluation Tech Rep No: 350). Santa Monica, CA: Rand Institute on Education and Training.

Kutlu, Ö., Doğan, C. H., & Karakaya, İ. (2010). *Öğrenci başarısının belirlenmesi performansa ve portfolyoya dayalı durum belirleme*. Ankara: Pegem Akademi Yayınları.

Lake, D. (1999). Helping Students to go SOLO: Teaching critical numeracy in the biological sciences. *Journal of Biological Education*, *33*(4), 191-198. doi:10.1080/00219266.1999.9655664

Leung, C. F. (2000). Assessment for learning: Using SOLO taxonomy to measure design performance of design & technology students. I*nternational Journal of Technology and Design Education*, *10*(2), 149-161. doi:10.1023/A:1008937007674

Lian, L. H., & Idris, N. (2006). Assessing algebraic solving ability of form four students. *International Electronic Journal of Mathematics Education*, *1*(1), 55-76. Retrieved from http://www.mathedujournal.com/dosyalar/a4.pdf

Lian, L. H., & Yew, W. T. (2012). Assessing algebraic solving ability: A theoretical framework. *International Education Studies*, *5*(6), 177-188. doi:10.5539/ies.v5n6p177

Linacre, J. M. (2014). *A user's guide to FACETS Rasch-model computer programs.* Retrieved from http://www.winsteps.com/a/facets-manual.pdf

Lucas, U., & Mladenovic, R. (2008). The identification of variation in students' understandings of disciplinary concepts: The application of the SOLO taxonomy within introductory accounting. *Higher Education*, *58*(2), 257-283. doi:10.1007/s10734-009-9218-9

McBee, M. M., & Barnes, L. L. B. (1998). The generalizability of a performance assessment measuring achievement in eight-grade mathematics. *Applied Measurement in Education*, *11*(2), 179-194. doi:10.1207/s15324818ame1102_4

McNamara, T. F. (1996). *Measuring second language performance*. London and New York: Longman.

Ministry of National Education. (2007). *Matematik öğretmen kılavuz kitabı*. Ankara: Devlet Kitapları Müdürlüğü.

Ministry of National Education. (2009). *İlköğretim matematik dersi 6-8. sınıflar öğretim programı*. Retrieved from http://ttkb.meb.gov.tr/program2.aspx

Ministry of National Education. (2013). *Temel eğitimden ortaöğretime geçişle ilgili sıkça sorulan sorular*. Retrieved from http://www.meb.gov.tr/duyurular/duyurular2013/bigb/tegitimdenoogretimegecis/MEB_SSS_20_09_2013.pdf

Mohd Nor, N., & Idris, N. (2010). Assessing students' informal inferential reasoning using SOLO taxonomy based framework. *Procedia Social and Behavioral Sciences*, *2*(2), 4805-4809. doi:10.1016/j.sbspro.2010.03.774

Mooney E. S. (2002). A framework for characterizing middle school students' statistical thinking, *Mathematical Thinking and Learning*, *4*(1), 23-63. doi:10.1207/S15327833MTL0401_2

Moore, B. B. (2009). *Consideration of rater effects and rater design via signal detection theory* (Unpublished doctoral dissertation). Columbia University, New York. Retrieved from http://search.proquest.com/docview/304862541

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation*, *7*(10), 71-81. Retrieved from http://pareonline.net/getvn.asp?v=7&n=10

Mulqueen, C., Baker D., & Dismukes, R. K. (2000, Nisan). *Using multifacet Rasch analysis to examine the effectiveness of rater training*. 15th Annual Conference for the Society for Industrial and Organizational Psychology (SIOP) konferansında sunulmuş bildiri, New Orleans. Retrieved from http://www.air.org/files/multifacet_Rasch.pdf

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and Measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applıed Measurement*, *5*(2), 189-227. Retrieved from http://jimelwood.net/students/grips/tables_figures/myford_wolfe_2004.pdf

National Assessment Governing Board. (2002). *Mathematics framework for the 2003 national assessment of educational progress.* Retrieved from http://academic.wsc.edu/faculty/jebauer1/mat645/framework_03.pdf

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. Reston, VA: Author. Retrieved from http://www.nctm.org/store/Products/Principles-and-Standards-for-School-Mathematics-(Book-and-E-Standards-CD)/

Özmantar, M. F., Bingölbali, E., & Akkoç, H. (2008, Mayıs). *İlköğretim sınıf öğretmenlerinin açık uçlu matematik soruları değerlendirme süreçleri*. Paper presented at VII. Ulusal Sınıf Öğretmenliği Eğitimi Sempozyumu, Çanakkale, Turkey. Retrieved from http://mimoza.marmara.edu.tr/~hakkoc/yayin2008_ozmantar_bingolbali_akkoc_usos.pdf

Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation*, *13*(4), 1-11. Retrieved from http://pareonline.net/getvn.asp?v=13&n=4

Purpura, J. E. (2004). *Assessing grammar*. Cambridge University Press.

Rembach L., & Dison, L. (2016). Transforming taxonomies into rubrics: Using SOLO in social science and inclusive education. *Perspectives in Education*, *34*(1), 68-83. Retrieved from http://scholar.ufs.ac.za:8080/xmlui/bitstream/handle/11660/3838/persed_v34_n1_a6.pdf?sequence=1&isAllowed=y

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher*, *94*(1), 31-37. Retrieved from http://www.peterliljedahl.com/wp-content/uploads/Myth-of-Objectivity.pdf

Romberg, T. E., & Wilson, L. D. (1992). Issues related to development of authentic assessment system for school mathematics. In T. A. Romberg (Ed.), *Reform in school mathematics and authentic assessment* (pp. 1-18). Albany: State University of New York Press.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413-428. doi:10.1037/0033-2909.88.2.413

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education. Retrieved from https://scale.stanford.edu/system/files/performance-assessment-era-standards-based-educational-accountability.pdf

Student Selection and Placement Center. (2013). *Açık uçlu sorularla deneme sınavının uygulanması*. Retrieved from http://www.osym.gov.tr/belge/1-19410/acik-uclu-sorularla-deneme-sinavinin-uygulanmasi-311201-.html

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*(3), 239-261. doi:10.1016/j.asw.2004.11.001

Tan, Ş. (2015). *Öğretimde ölçme ve değerlendirme KPSS el kitabı*. Ankara: Pegem Akademi Yayıncılık.

Tekin, H. (2009). *Eğitimde ölçme ve değerlendirme*. Ankara: Yargı Yayınevi.

Toffoli, S. F. L., Andrade, D. F., & Bornia, A. C. (2016). Evaluation of open items using the many-facet Rasch model. *Journal of Applied Statistics*, *43*(2), 299-316, doi:10.1080/02664763.2015.1049938

Walker, E. R., Engelhard, G., & Thompson, N. J. (2012). Using Rasch measurement theory to assess three depression scales among adults with epilepsy. *Seizure*, *21*(6), 437-443. doi:10.1016/j.seizure.2012.04.009

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17*(3), 150-173. doi:10.1016/j.asw.2011.12.001

Woodward, J., Monroe, K., & Baxter, J. (2001). Enhancing student achievement on performance assessments in mathematics. *Learning Disability Quarterly, 24*(1), 33-46. doi:10.2307/1511294

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, *8*(3), 370.

Yazıcı, N. (2013). *Başarının ölçülmesinde SOLO taksonomiye dayalı hazırlanan rubrik kullanımının etkisinin karşılaştırmalı olarak incelenmesi* (Unpublished master's thesis). Kahramanmaraş Sütçü İmam University, Institute of Social Science, Kahramanmaraş, Turkey.

Zhu, X. (2009). *Assessing fit of item response models for performance assessments using bayesian analysis* (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh, ABD. Retrieved from http://d-scholarship.pitt.edu/10162/1/XiaowenZhu_ETD2009_Final.pdf

Appendix 1

An Open-ended Mathematical Question

If p is a real number, which is larger, $2p$ or $p+6$?

A Standard Rubric

Rating Criteria	
3 points ***Excellent***	The problem is perfectly understood. -To find which of the two mathematical expressions is larger, the student established relationships of equality and inequality between $2p$ and $p+6$ and found the correct answer. The answer is below. It is a clear, understandable and exemplary solution. $2p>p+6$ is correct when $p>6$. $2p=p+6$ is correct when $p=6$. $p+6>2p$ is correct when $p<6$.
2 points ***Good***	The problem is generally understood. -The solution is mainly correct except for minor errors. The student used the relations, $2p>p+6$, $2p=p+6$ and $p+6>2p$. However, the student failed to complete the task due to minor calculation errors or other unclear reasons. -The student found the correct answer, which is; the two expressions are equal when $p=6$ and $2p$ is larger when $p>6$ and $p+6$ is larger when $p<6$. However, the explanation of how the student solved the problem is insufficient.
1 point ***Acceptable***	The problem was partially understood. -The student started to solve the problem with correct strategies such as establishing equalities or inequalities between $2p$ and $p+6$. However, they failed to complete the problem. -The student was only able to start with the correct strategy. There are major errors in the operations done by the student.
0 points ***Inadequate***	The problem was not understood. -The student wrote down notes such as "The values of $2p$ and $p+6$ are unknown, so it is not possible to determine which is larger." -The student did not do any operations to determine whether $2p$ or $p+6$ was larger. -The student wrote, "We are asked to find whether $2p$ or $p+6$ is larger," which merely restates the problem. -The student used an incorrect strategy to find which expression is larger.

A Rubric Based on the SOLO Taxonomy

Rating Criteria	
3 points ***Relational***	The student is capable of appointing $p=6$ as the critical value and assumes that there are different conditions for p being above or below 6. The student found the correct answer: $2p=p+6$ when $p=6$; $2p>p+6$ when $p>6$, and $p+6>2p$ is correct when $p<6$.
2 points ***Multistructural***	The student is aware that p is a variable. The student tried to solve the problem by appointing multiple values to the p variable. The student is able to make an interpretation by appointing different values to p yet fails to consider all possible situations. In particular, the student is not aware that the two statements are equal when $p=6$, and that the results would differ with values larger and smaller than 6. The student may write, "$p+6$ is larger when $p=2$, and $2p$ is larger when $p=10$. So whether $2p$ or $p+6$ is larger depends on the circumstances."
1 point ***Unistructural***	The student attempted to solve the problem by assigning a single value to p. The student is aware of the concept of variables. However, the student used a one-dimensional approach to the problem. The student's response to the problem may be: "When $p=3$, $2p=6$ and $p+6=9$. Therefore, $p+6$ is larger than $2p$."
0 points ***Prestructural***	The student has difficulty understanding the problem. The student's responses are not relevant. Since the student does not have any idea about the concept of variables, they may add dissimilar terms such as $p+6=7$ or assign different values to p in $2p$ and $p+6$.